Proc. Natl. Acad. Sci. USA
Vol. 90, pp. 600–604, January 1993
Genetics

Genome mapping by nonrandom anchoring: A discrete theoretical analysis

M. Q. ZHANG AND T. G. MARR

Cold Spring Harbor Laboratory, P.O. Box 100, Cold Spring Harbor, NY 11724

Communicated by Joel L. Lebowitz, October 5, 1992 (received for review May 5, 1992)

ABSTRACT **As part of our effort to construct a physical map of the genome of the fission yeast *Schizosaccharomyces pombe*, we have made theoretical predictions for the progress expected, as measured by the expected length fraction of island coverage and by the expected properties of the anchored islands such as the number and the size of islands. Our experimental strategy is to construct a random clone library and screen the library for clones having unique sequence at both ends. This scheme is essentially the same as the clone-limited double sequence-tagged-site selection scheme which was used in a computer simulation by Palazzolo *et al.* [Palazzolo, M. J., Sawyer, S. A., Martin, C. H., Smoller, D. A. & Hartl, D. L. (1991) *Proc. Natl. Acad. Sci. USA* 88, 8034–8038]. Both simulation and ongoing experiments in our laboratory have shown that the nonrandom anchoring method is far superior to random anchoring. In this paper, we propose a theoretical model to explain the simulated data and the experimental data.**

One of the short-term goals of the Human Genome Project is to produce libraries of largely contiguous, ordered sets of molecular clones for use in sequencing and gene mapping projects. This is planned to be done for the genomes of human and several model organisms such as yeasts, nematode, fruit fly, and mouse. Given a library of random clones, there are two major ordering schemes for building contigs (collections of overlapping clones). One is known as "fingerprinting" (1–4); the other is known as STS (sequence-tagged site) content-detection, or "anchoring" (5–7). Two clones from the random clone library are considered linked if they share common fingerprints or common anchor sequences.

Theoretical analysis for the case where the anchors are chosen randomly from the genome and independently of the clones has been done by a number of authors (refs. 8–12 and unpublished work). In this paper, we propose a theoretical model which provides a reasonably close quantitative account of a nonrandom anchoring method that has been implemented in our laboratory (unpublished data). The mapping scheme may be described as follows: (*i*) build a random clone library; (*ii*) take an arbitrary clone which has not been anchored and has unique sequences at both ends; (*iii*) make a pair of probes complementary to these two ends; (*iv*) use this pair to probe the library in order to anchor more clones; (*v*) repeat steps *ii–iv* until almost all the unanchored clones have been exhausted; (*vi*) then start chromosome-walking from the edges of each contig built and try to fill up the gaps. In the end, a number of contigs spanned by the anchored clones is obtained. To determine the physical length of a contig, one has to measure the lengths of its spanning clones by digesting with restriction enzymes. Finally, one may use some known gene probes or other STS probes to anchor each contig back onto the chromosome. Steps *i–v* of this contig-building strategy are exactly the ones simulated by Palazzolo *et al.* (14) in their computer experiment if we interpret their STSs as our probes (we do not have to know the actual sequence). They realized the potential value of this mapping scheme after they had compared it with the simulations of random selection of probes and single-ended clone-limited methods.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

An Approximate Theoretical Model

The exact mathematical description of the scheme is extremely complicated. The difficulty lies in the fact that the probe positions and the clone positions are highly correlated, so the distribution of the probes is conditioned upon the distribution of the clones. We propose an approximate model which is capable of explaining most of the relevant features of the mapping strategy. The idea is to separate clones into two groups: those used for making probes, called "probing clones"; the rest are called "target clones." We assume the target clones are distributed randomly and uniformly over the genome; independent of the target clones, the probing clones are also distributed randomly and uniformly except that they cannot overlap among themselves. The total number of probing clones and target clones must be equal to the total number of the clones in the library (the end effect on the genome size is also ignored). Our approximation is attainable because of the independence assumption; otherwise the total clone distribution is uniform, so knowing the positions of the target clones biases the distribution of the probing clones and vice versa. This approximation is expected to be a good one if the number of target clones is large, as is typically the case when constructing a redundant genomic library. Another subtle, but key, assumption is that we assume simultaneous probing, whereas in practice, probing is sequential. This latter assumption affects mainly the "stopping time" (i.e., when probing clones get "jammed") and will have a significant effect on the results only if the underlying library is not random (for example, if one is using a selective cloning technique). With simultaneous probing, jamming occurs near the close-packing threshold, which is relatively near the maximum density point (the maximum density of probing clones is 1); with sequential probing, jamming occurs well before the close-packing threshold (for a large number of clones, the expected critical density of probing clones is $\rho \approx 0.7476$).*

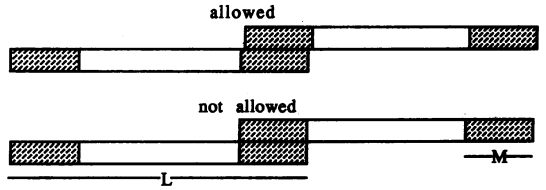
Our general setup is similar to the one in our paper to be published elsewhere. Let the genome size be G, the clone size be L, and the size of each probe end be M, all measured in units of base pairs. Let the probing clone number (which is the same as the left probe number or the right probe number), target clone number, and the total clone number be N_s, N_c and N, respectively; so that $N_s + N_c = N$. We define $c = N_c/G$, $s = N_s/G$, $a = NL/G$, $b = 2sL$, and $t = M/L$.

Before we start to calculate the interesting quantities, we first discuss some important properties of the probing clones.

Abbreviation: STS, sequence-tagged site.

*This is a famous parking problem or a random sequential absorption problem; see for instance ref. 15. The exact value is given by the integral $\int_0^\infty \exp[-2\int_0^y x^{-1}(1 - e^{-x})dx]dy$.

Hard-Rod Fluid Properties of Probing Clones. Since probing clones cannot overlap (this assumption implements the key fact that only an unanchored clone can be used to make a pair of probes from its ends), they behave like one-dimensional fluids made of hard rods. The fundamental property of hard-rod fluids is the exclusion that a rod of length D starting at position x (i.e., its left end is at x) will prevent the positions $x - D + 1, x - D + 2, \ldots, x + D - 1$ from being occupied by a left end of any other rod; in other words, the left ends of any two consecutive rods have to be at least D distance away from one another. In the case of probing clones, the quantity, $(L - M + 1)$ is the equivalent of the hard-rod length, D.



From this basic property, one can easily derive the following properties which will be used in later formulations. We list the useful properties below without proofs (one can find the derivations in ref. 16). Assume the density of the rod at position x is $n(x)$, which is the probability that x is occupied by the left end of a rod (we use the left end throughout; any other point of a rod can also be taken as the reference point just as well). Then the probability that x is not occupied by the left end of a rod, which we denoted by $n(\bar{x})$, is equal to $1 - n(x)$. For the uniform distribution, they are just constants s, $1 - s$. If we define

$$m(x, y) \equiv 1 - \sum_{j=1}^{y} n(x - j + 1)$$

and its uniform version

$$m(y) \equiv 1 - ys,$$

then

1. The probability that the positions $x - y + 1, x - y + 2, \ldots, x$ are not occupied is

$$n(\overline{x - y + 1}, \overline{x - y + 2}, \ldots, \bar{x}) = \begin{cases} m(x, y) & \text{if } y \le D, \\ \dfrac{\prod_{i=1}^{y-D+1} m(x - i + 1, D)}{\prod_{i=1}^{y-D} m(x - i, D - 1)} & \text{if } y > D. \end{cases}$$

If the distribution is uniform, this probability reduces to

$$n(\overline{x - y + 1}, \overline{x - y + 2}, \ldots, \bar{x}) = \begin{cases} m(y) & \text{if } y \le D, \\ \dfrac{m(D)^{y-D+1}}{m(D-1)^{y-D}} & \text{if } y > D. \end{cases} \quad [1]$$

2. The probability that the positions $x - y + 1, x - y + 2, \ldots, x - 1$ are not occupied and x is occupied is

$$n(\overline{x - y + 1}, \overline{x - y + 2}, \ldots, \overline{x - 1}, x) = \begin{cases} n(x) & \text{if } y \le D, \\ n(x) \prod_{i=1}^{y-D} \dfrac{m(x - i, D)}{m(x - i, D - 1)} & \text{if } y > D. \end{cases} \quad [2]$$

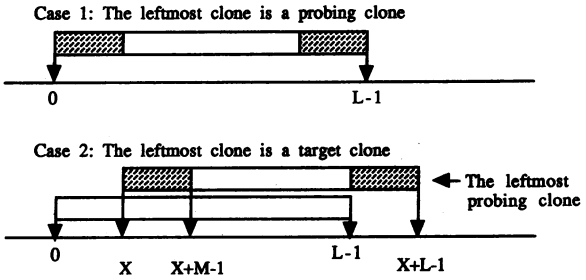
If the distribution is uniform, this probability reduces to

$$n(\overline{x - y + 1}, \overline{x - y + 2}, \ldots, \overline{x - 1}, x) = \begin{cases} s & \text{if } y \le D, \\ s\left(\dfrac{m(D)}{m(D - 1)}\right)^{y-D} & \text{if } y > D. \end{cases} \quad [3]$$

Equipped with these probabilities, we now calculate the expectations of interest.

Expected Number of Islands. When we say "island," we always mean "anchored island," which includes singly anchored clones, probing clones, and contigs. In a paper to be published elsewhere, we will derive discrete formulations for the random anchoring scheme.

First, we calculate the expected number of islands. For simplicity of notation, we denote $\bar{c} \equiv 1 - c$ and $\bar{s} \equiv 1 - s$. We assume that $L > M$ and that a target clone is anchored if at least one probe end is completely contained within the target clone. The number of islands is of course equal to the number of the left ends of islands. The event that there is a left end of an island starting (from the left to right) at some arbitrary position 0 is characterized by the following: the leftmost clone of the island is a probing clone or the leftmost clone is an anchored target clone.



The probability p that an island starts at position 0 is

$$\bar{c}^{L-M}s + c\sum_{x=1}^{L-M} \bar{c}^{L-M-x} n(\overline{M - L}, \overline{M - L + 1}, \ldots, \overline{x - 1}, x)$$

$$= \bar{c}^{D-1}s + cs\,\frac{\bar{c}^{D-1} - \left(\dfrac{1 - Ds}{1 - (D - 1)s}\right)^{D-1}}{\bar{c} - \dfrac{1 - Ds}{1 - (D - 1)s}}\left(\frac{1 - Ds}{1 - (D - 1)s}\right), \quad [4]$$

where we have made use of Eq. 3. This probability depends only on the difference $D = L - M + 1$ (unpublished work).

The expected number of islands, N_{island}, is given by Gp.

The above is the general expression for the model. If the lengths of the genome, clones, and probes are much larger than 1 base pair, we can approximate the exact result by its continuous limit:

$$c, s \to 0, \text{ but } cL \to a - \frac{b}{2} \equiv \alpha, sL \to b/2 \equiv \beta.$$

We then have

$$N_{\text{island}} \to N_s e^{-(1-t)\alpha} + N_s\alpha\,\frac{e^{-\frac{(1-t)\beta}{1-(1-t)\beta}} - e^{-(1-t)\alpha}}{(1 - t)\alpha - \dfrac{(1 - t)\beta}{1 - (1 - t)\beta}}. \quad [5]$$

In most applications, we can simply set $t = 0$ because the size of the probes is usually much smaller than that of the clones. We obtain

$$N_{\text{island}} = N_s e^{-\alpha} + N_s \alpha \frac{e^{-\frac{\beta}{1-\beta}} - e^{-\alpha}}{\alpha - \frac{\beta}{1-\beta}}. \quad [6]$$

In our approximate model, the density of probing clones, β, varies from 0 to 1 (the close-packing limit). As noted before, we had to make the assumption of simultaneous probing, while recognizing that, in practice, sequential probing occurs. This means we may only apply our formulas for β less than the critical density, 0.75. This is equivalent to allowing b (the density of probes) to vary from 0 to 1.5.

To compare with the random anchoring scheme, we have plotted Eq. 6 and the corresponding random anchoring results in Fig. 1. The acceleration of the nonrandom scheme is significant. For example, with a 5-fold coverage library ($a = 5$) at the jamming density ($b = 1.5$), the nonrandom scheme would bring the island number down to $0.15G/L$, whereas the random scheme would have to set $b = 5.4$ in order to get the same results. For a genome of size $G = 14$ megabase pairs (Mb) and a clone length $L = 40$ kilobase pairs (kb), 1,365 additional probing experiments would have to be done using the random scheme.

The Expected Coverage of the Genome and the Expected Length of an Island

Second, we calculate the expected fraction of genome length covered by islands, which should serve as a measure of the progress of a mapping project. We begin with calculating the probability that a position is not covered by any islands.

For an arbitrary position 0, the probability that 0 is not covered by any clones (neither probing clones nor target clones) is

$$\bar{c}^L n(\overline{-L+1}, \overline{-L+2}, \ldots, \bar{0}) = \bar{c}^L \frac{(1-Ds)^M}{(1-(D-1)s)^{M-1}}.$$

The probability q that 0 is covered by one unanchored clone is

$$Lc\bar{c}^{L-1} n(\overline{M-L}, \overline{M-L+1}, \ldots, \overline{L-M}) = Lc\bar{c}^{L-1} \frac{(1-Ds)^D}{(1-(D-1)s)^{D-1}}.$$

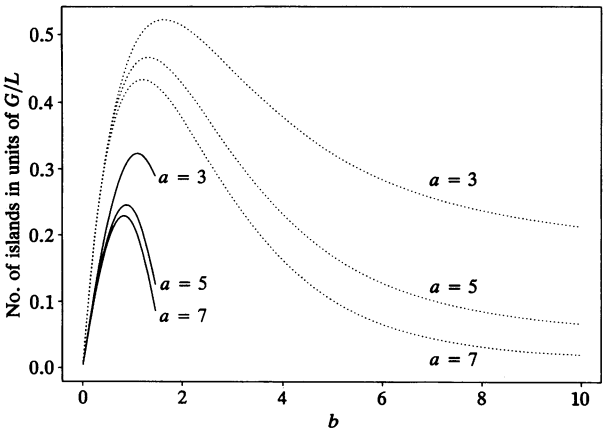
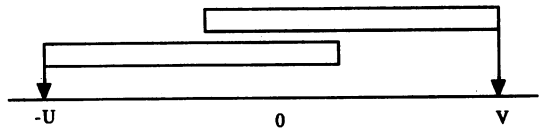


FIG. 1. The expected number of islands for the accelerated scheme (solid lines) and the random scheme (dotted lines).

To calculate q, we assume that the left end of the leftmost target clone covering 0 is at $-u$ (which means $1 \le u \le L-1$) and the right end of the rightmost target clone covering 0 is at v (which means $-u + L \le v \le L - 1$).



Therefore

$$q = \sum_{u=1}^{L-1} \sum_{v=-u+L}^{L-1} c^2 \bar{c}^{2L-2-u-v} \cdot n(\overline{-u+M-L}, \overline{-u+M-L+1}, \ldots, \overline{v+1-M}),$$

$$= c^2 \frac{1-(D-1)s}{\bar{c}-\gamma} \gamma^D \left[(L-1)\bar{c}^{L-1} - \gamma \frac{\bar{c}^{L-1} - \gamma^{L-1}}{\bar{c}-\gamma} \right],$$

where

$$\gamma \equiv \frac{1-Ds}{1-(D-1)s}.$$

Summing up these probabilities for the three exclusive events, we find the probability that an arbitrary site is not covered by an island is

$$r_0 = \bar{c}^L \frac{(1-Ds)^M}{(1-(D-1)s)^{M-1}} + Lc\bar{c}^{L-1} \frac{(1-Ds)^D}{(1-(D-1)s)^{D-1}} + q. \quad [7]$$

Again, if we are interested only in the special case of the continuous limit, we get

$$r_0 \to [1-(1-t)\beta] e^{-\alpha - \frac{t\beta}{1-(1-t)\beta}} + \alpha [1-(1-t)\beta] e^{-\alpha - \frac{(1-t)\beta}{1-(1-t)\beta}} + \alpha^2 [1-(1-t)\beta] e^{-\frac{(1-t)\beta}{1-(1-t)\beta}} \times \left[-\frac{e^{-\alpha}}{\alpha - \frac{(1-t)\beta}{1-(1-t)\beta}} + \frac{e^{-\frac{(1-t)\beta}{1-(1-t)\beta}} - e^{-\alpha}}{\left(\alpha - \frac{(1-t)\beta}{1-(1-t)\beta}\right)^2} \right].$$

If probe length is negligible, we get

$$r_0 \to (1-\beta)e^{-\alpha} + \alpha(1-\beta)e^{-\alpha - \frac{\beta}{1-\beta}} + \alpha^2(1-\beta)e^{-\frac{\beta}{1-\beta}} \left[-\frac{e^{-\alpha}}{\alpha - \frac{\beta}{1-\beta}} + \frac{e^{-\frac{\beta}{1-\beta}} - e^{-\alpha}}{\left(\alpha - \frac{\beta}{1-\beta}\right)^2} \right]. \quad [8]$$

The fraction of the genome length covered by islands, μ, is of course given by $1 - r_0$. This is plotted in Fig. 2 for $t = 0$. Comparing with the corresponding random scheme shown in the same figure, we can clearly see the acceleration of the rate toward the completion of the genome by the nonrandom scheme.

To measure l, the expected size of an island, we use the approximate formula (assuming that the amount of cryptic overlap between islands is negligible)

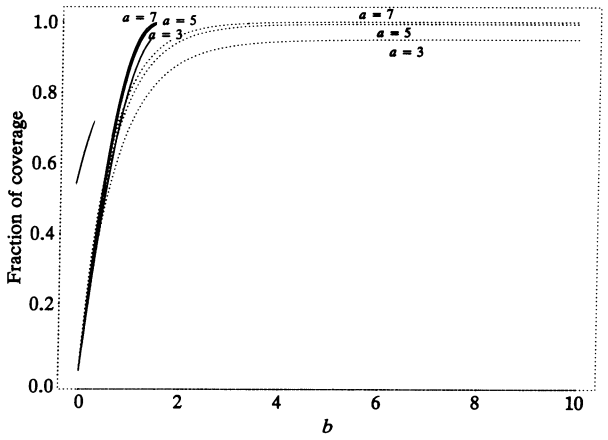


FIG. 2. The fraction of genome coverage for accelerated scheme (solid lines) and the random scheme (dotted lines).

$$l \approx \frac{\text{Total length covered by islands}}{\text{Total number of islands}} = \frac{G(1-r_0)}{N_{\text{island}}},$$

or in units of the clone length L,

$$l = \frac{1-r_0}{Lp}. \qquad [9]$$

In Fig. 3, the expected island size is plotted against b, which is proportional to the total number of probes. Expected island size grows much faster when the nonrandom anchoring method is used than if one were to use random anchoring.

Comparison of the Approximate Model with Computer Simulations and with Experiments

For the double-end clone-limited STS assay, Palazzolo *et al.* (14) did a computer simulation. Their genome size was set at 100 Mb and the mapping clone insert at 100 kb, and the library contained a total of 5000 clones which gave a 5-fold coverage (i.e., $a = 5$). We have plotted our theoretical curves (solid lines) of the percent genome coverage (Fig. 4*A*), the number of islands (Fig. 4*B*), and the average island size (Fig. 4*C*) against $b = 2\beta$ for the fixed $a = \alpha + \beta = 5$. We have superimposed the simulation data (squares) (from figure 1 of ref. 1) together with the corresponding random anchoring curves (dashed lines). The experimental data (unpublished) from the *Schizosaccharomyces pombe* genome mapping research at our laboratory are plotted (crosses) for the number of islands (Fig. 4*B*) only; because restriction mapping has not been completed, the information on the average size of islands is not yet available. A number of points based on the comparison of these curves is in order.

(*i*) We have terminated the theoretical curve at $b = 1.5$, which corresponds to the critical jamming density of probing clones, as explained earlier. The general agreement between the theory and the simulations is remarkably good. The slightly earlier termination (around $b = 1.4$) of the simulation data is presumably due to the finite coverage of the library (not an infinite ensemble).

(*ii*) The best agreement between the theory and the simulation data is in the small b region, where the number of the

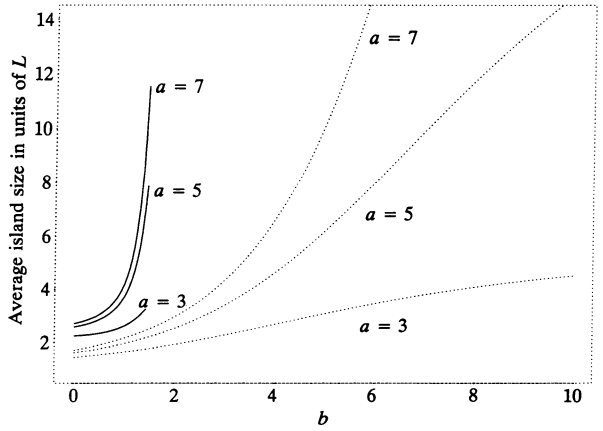


FIG. 3. The average size of an island for the accelerated scheme (solid lines) and the random scheme (dotted lines).

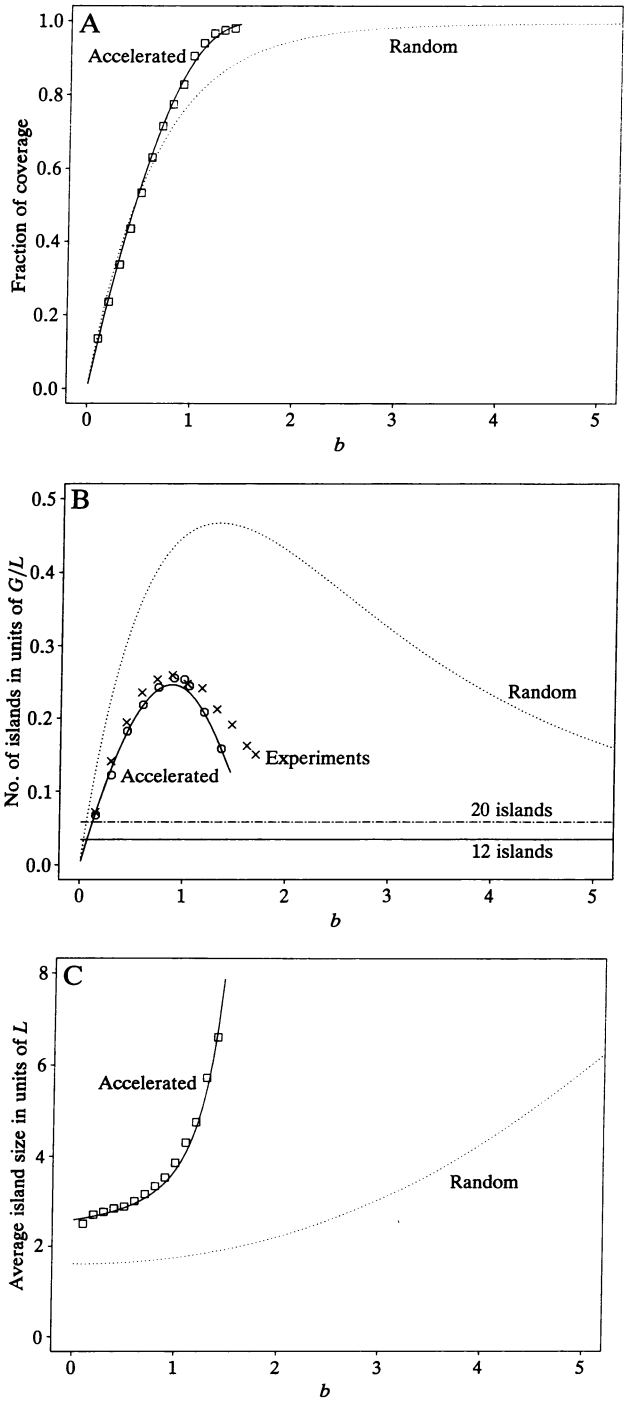


FIG. 4. Comparison of the theory (circles), the computer simulations (squares) and the experiments (crosses) for $a = 5$. The dotted curve represents the corresponding random scheme.

probing clones is small, which does not bias the target clone distribution too much. When b gets relatively large, the bias effect starts to show up. This effect tends to cause the number of islands to be less than would be predicted using the true model, because, in the true model, the target clones and probing clones are correlated so that knowing there is a probing clone should lower the chance of anchoring a target clone, if the total clone distribution is uniform. This translates to slightly greater island length and larger genome fraction of coverage than the true model. It seems that these discrepancies are extremely small according to the available data. This indicates that our approximate model is, though simple, an adequate one for most practical purposes.

(*iii*) The so-called "Lander–Waterman hump," which indicates the joining of islands, occurs at much lower value of island number when compared with the random anchoring scheme. This shows that the potential speed of the nonrandom scheme is much faster. We shall not elaborate more on the differences between the two schemes; the original paper (14) gives a fairly detailed account of the differences.

(*iv*) In our *S. pombe* genome mapping project, a library of 1700 (N) cosmid clone inserts of size about 40 kb (L) was used for building high-efficiency cosmid vectors containing bacteriophage T3/T7 promoter sequences flanking the cloning site to allow for the synthesis of end-specific probes (13). This gives a 5-fold coverage ($a = 5$), because the genome size was estimated to be about 13.8 Mb (G) excluding the telomere, centromere, and two repeat regions. The T3/T7 probes made from the two ends of mapping clones were of 200–1000 base pairs, which could be neglected so that we may apply the continuous formula with $t = 0$. We plot every relevant quantity against $b = 2N_sL/G$, as it is proportional to the total number of probes used and therefore is in turn proportional to the number of experiments done. In the actual experiments in our laboratory, some known genes were also used as probes in order to facilitate the mapping of contigs to the chromosomes. Since the redundancy due to these gene probes could obscure the underlying principle, we plot only the number of islands anchored by the cosmid clone probes vs. the number of the cosmid clone probes.

(*v*) To compare with the actual experiments in our laboratory, we have to bear in mind that, due to technical reasons, the actual procedure was a bit different from the ideal protocol. For example, instead of probing sequentially one by one, the probes were made and used in groups of 8 or 16 (from the ends of either four or eight clone inserts) because of the capacity of the equipment and the work limit per day. In addition, one group of probes was being used in hybridizations while another group of probes was being prepared. All of these factors would slightly increase the redundancy because probes could overlap among themselves, which would make the "hard rods" softer. This should translate into a slightly higher number of islands, shorter islands, and lower fractional coverage of the genome at the same b (i.e., at the same number of experiments). And indeed, this is so, as can be seen in Fig. 4*B*.

(*vi*) Many of the progressive deviations of the experimental data from the theoretical curve were caused by the situation where some ends of the anchored clones were used to make new probes, which created further redundancies. Ideally, one would like to use unanchored clones to make probes until the jamming threshold is about to be reached, and then use the far ends of the islands to make probes and to start chromosome walking in order to reduce the redundancy. In practice, some unanchored clones were not detected because of the weak signals; they were subsequently chosen to make the end probes, and this also increased the redundancy. In addition, as the jamming threshold was approached, most of the islands had been joined, further progress became very slow, and chromosome walking was started at about $b = 1.3$. Since our theory cannot be used beyond the critical jamming ($b = 1.5$), we could not predict the path of the walking experiments. Theoretically, the walking could begin at $b = 1.5$. Because of the "critical slowing down," which depends very much on the quality of the library and the actual experimental procedure, we suggest that walking begin when b is 1.3–1.4, concomitantly with the phasing out of anchoring.

(*vii*) If all of the possible ends of the islands are tried as probes, regardless of how many gene probes or any other probes are used, the number of islands will approach the number of actual islands, which is a property of the library itself. The expected value is given by ae^{-a} and is equal to $0.034G/L$ (the solid horizontal line in Fig. 4*B*) for $a = 5$, which corresponds to about 12 islands with our experimental parameters.

(*viii*) To compare the predicted average length of an island with the experimental results, we have to use the physical map created from restriction enzyme digestions of the spanning clones of each island to measure the island length. We have already been able to compare some preliminary data with the theory and to make suggestions concerning some possible errors that may have occurred in the experiments.

The value of the theoretical results is not limited to explaining the experimental data; we think it is far more important to use the theory in designing a new mapping project or in comparing different mapping schemes in order to select an optimal one, or in guiding the experiments after a project starts. In the hybridization experiments, the number of islands could tell how many probings one has to do, whether there have been too many false positives or false negatives, when one should start chromosome walking, and what the end product might be. In the digestion experiments, the expected island size could help detect whether an erroneous restriction map had been made (after taking into account the trimming effect by estimating the average restriction fragment length). If the measured length is too long, it may be due to some partial digestions; if it is too short, it may be due to missing (nearly identical) fragments.

We thank T. Mizukami and W. I. Chang for providing the experimental data and helpful discussions in a comparable form. This work was supported by National Institutes of Health Grant 1R01 HG00203-01A1 and Department of Energy Grant DE-FG02-91ER61190 to T.G.M. T.G.M. was also partially supported by National Institutes of Health Grant R01 HG00017-02 to D. Beach.

1. Olson, M. V., Dutchik, E., Graham, M. Y., Brodeur, G. M., Helms, C., Frank, M., MacCollin, M., Scheinman, R. & Frank, T. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 7826–7830.
2. Coulson, A., Sulston, J., Brenner, S. & Karn, J. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 7821–7825.
3. Kohara, Y., Akiyama, K. & Isono, K. (1987) *Cell* **50**, 495–580.
4. Stallings, R. L., Torney, D. C., Hildebrand, C. E., Longmire, J. L., Deaven, L. L., Jett, J. H., Doggett, N. A. & Moyzis, R. K. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 6218–6222.
5. Green, E. D. & Olson, M. V. (1990) *Science* **250**, 94–98.
6. Williams, J. G. K., Kubelik, A. R., Livak, K. J., Rafalski, J. A. & Tingey, S. V. (1990) *Nucleic Acids Res.* **18**, 6531–6535.
7. Welsh, J. & McClelland, M. (1990) *Nucleic Acids Res.* **18**, 7213–7218.
8. Lander, E. S. & Waterman, M. S. (1988) *Genomics* **2**, 231–239.
9. Barillot, E., Dausset, J. & Cohen, D. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 3917–3921.
10. Torney, D. (1991) *J. Mol. Biol.* **217**, 259–264.
11. Ewens, W. J., Bell, C. J., Donnelly, P. J., Dunn, P., Matallana, E. & Ecker, J. R. (1991) *Genomics* **11**, 799–805.
12. Arratia, R., Lander, E. S., Tavaré, S. & Waterman, M. S. (1991) *Genomics* **11**, 806–827.
13. Evans, G. A., Lewis, K. & Rothenberg, B. E. (1989) *Gene* **79**, 9–20.
14. Palazzolo, M. J., Sawyer, S. A., Martin, C. H., Smoller, D. A. & Hartl, D. L. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 8034–8038.
15. González, J. J., Hemmer, P. C. & Hoye, J. S. (1974) *Chem. Phys.* **3**, 228–238.
16. Zhang, M. Q. (1991) *J. Statist. Phys.* **63**, 1191–1202.